ABSTRACT

This paper provides a formal statistical analysis of three important aspects of parameterized retrieval models: estimation, sensitivity, and generalization. Since all parameterized models, even those based on heuristics, have inherent uncertainty, we investigate these aspects within a robust statistical framework. The results of experiments carried out over a wide range of popular retrieval models and data sets show that using a maximum posterior estimate yields good effectiveness for stable distributions, whereas Bayesian techniques are more appropriate for unstable distributions. We also show that language modeling-based models are typically less sensitive and generalize better within and across collections than BM25 variants.

Categories and Subject Descriptors
H.3.3 [Information Storage and Retrieval]: Information Search and Retrieval

General Terms
Algorithms, Experimentation, Theory

Keywords
Retrieval models, estimation, sensitivity, generalization

1. INTRODUCTION

In information retrieval, a majority of the most commonly used retrieval models are parameterized models. A parameterized model is one whose scoring function has one or more tunable parameters. In these kinds of models, there are several important issues that have not received a great deal of attention, but are critical for developing a deeper understanding of such models.

First, there is the issue of choosing a parameter setting that effectively ranks documents according to some metric. How does one choose a parameter setting that leads to good effectiveness? There are currently several options. One could use the de facto method of tuning for some metric or use one of the various ‘default’ parameter settings that have been proposed. Unfortunately, there is little empirical or theoretical evidence to back the use of any one method. In this work, we investigate several selection techniques and weigh the pros and cons of each.

Next, parameter sensitivity is an important issue because it tells us how stable or erratic our models are to slight variations in the parameter setting. It may be difficult to choose an effective parameter setting for a sensitive model. Here, we examine connections between parameter sensitivity and estimation.

Finally, generalization may be the most important issue surrounding parameterized retrieval models. Given a parameter setting that is trained on some collection, how well will it do on a new set of topics on the same collection? How well will it do on a completely different document collection? The answers to these questions are not only important in TREC-style evaluations, but also in practical systems. When developing a product that will be used across a wide range of collections, it is important to choose a model that has good cross-collection generalization properties, because it is likely a single parameter setting will be used across all collections.

In the remainder of this paper we take a novel statistical approach in addressing each of these issues. Our work formalizes and extends the current understanding of the relationship between parameters and models. In addition to providing deeper insight into existing models, the analytical tools proposed here can also be used to develop better models in the future.

The rest of this paper is laid out as follows. First, Section 2 discusses related work. Next, Section 3 addresses the parameter selection problem and proposes a number of solutions to it. Then, Sections 4 and 5 tackle parameter sensitivity and generalization issues, respectively. Finally, experimental results are provided in Section 6 and are followed by conclusions in Section 7.
2. RELATED WORK

Up until this point, there has been little work in the information retrieval literature that has looked at parameter estimation, sensitivity and generalization in any substantial detail.

Parameter selection is typically done by choosing the 'default' model parameters or by optimizing for some metric. No analysis has been done to determine if there are, perhaps, more appropriate methods.

Despite the lack of substantial theoretical results for parameter selection, there have been a number of interesting results in the area of parameter sensitivity [3, 4, 12, 13]. Zhai and Lafferty [12] were the first to propose looking at the sensitivity of smoothing parameters in the language modeling framework to retrieval. More recently, Fang and Zhai [3, 4] studied how to formalize retrieval heuristics. They proposed an axiomatic approach to generating retrieval functions based on a set of retrieval parameter constraints. This approach yields models that are less sensitive than existing models. All of this work, however, qualitatively analyzes sensitivity via the use of plots. Little is done to address the issue quantitatively.

Finally, there has been little done to investigate how well current retrieval models generalize in different scenarios and if there is actually any connection between parameter sensitivity and generalization. In this work we hope to fill in some of these holes, extend the current understanding, and stimulate further research into these issues.

3. PARAMETER ESTIMATION

In information retrieval, the goal of parameter estimation is to choose a parameter setting that yields the most effective model possible. The definition of 'effective' varies across retrieval scenarios and tasks. For ad hoc retrieval, this may mean choosing a parameter setting that produces good mean average precision numbers. In known item search and question answering, this means optimizing a metric such as mean reciprocal rank.

Most of the analysis in this paper assumes that some form of training data is available. This data can come in the form of explicit relevance judgments or web clickthrough data, for example [5]. Later in this paper we briefly examine how well models perform in a completely unsupervised environment.

In the remainder of this section we explore the supervised parameter estimation problem in detail. Since uncertainty is inherent in all parameterized retrieval models, even those based on heuristics, the problem is best motivated and studied in a statistical framework. The framework we choose to work in is based on the parameter posterior distribution. As we will show, such a distribution can easily be estimated using a simple bootstrap technique.

3.1 Problem Statement

Given a model, parameterized by \( \theta \), let \( \mathcal{M}_\theta \) be the parameter space \(^1\) and \( m(\theta', T) \) be the value of the effectiveness metric evaluated at \( \theta' \) with regard to training data \( T \). Furthermore, let \( P(\theta'|T) \) be the likelihood that parameter setting \( \theta' \) is the optimal parameter setting given the training data.

\(^2\)This is the posterior distribution over optimal parameter settings after observing the training data.

In light of this definition, we see that there are two distinct ways of representing the effectiveness characteristics of the model. Given a fixed training set, both \( m(\cdot, T) \) and \( P(\cdot, T) \) can be thought of as functions over \( \mathcal{M}_\theta \), the model’s parameter space. Therefore, we can view things either from a geometric point of view or from a statistical point of view.

Figure 1 shows the correspondence between the metric surface, \( m(\theta', T) \), and the estimated posterior, \( P(\theta'|T) \), across a range of collections. We now look at how these dual representations can be used for parameter selection.

3.2 Point Estimates

The most common approach to parameter selection involves choosing a single parameter setting for use on all queries. Parameters selected in this way are known as point estimates. In this scheme, a new parameter is typically selected for each collection, if possible. The underlying assumption of such a selection policy is that the model itself can account for the variance in the different query types and, therefore, the parameters are collection dependent, but not query dependent. Although it is not clear how valid this assumption is, this is by far the most widely used approach. There have been attempts to select parameters on a query by query basis, which has been met with some success [9].

We now describe two point estimation techniques based on the metric surface and the posterior distribution, respectively.

3.2.1 Maximum Metric

The first point estimate technique is based on the metric surface. In this approach, the parameter that maximizes the metric over the parameter space is selected. That is,

\[ \hat{\theta} = \arg \max_{\theta'} m(\theta', T) \]

where \( \hat{\theta} \) is our estimate. The rationale behind this choice is that the parameter selected will produce good retrieval effectiveness on a test set as long as the maximum of the underlying metric surface of the unseen data takes on a similar value.

3.2.2 Maximum A Posteriori

The other point estimate technique is based on the posterior distribution. The parameter that maximizes the posterior is selected. Or, more formally,

\[ \hat{\theta} = \arg \max_{\theta'} P(\theta'|T) \]

which is known as the maximum a posteriori estimate. The motivation behind this approach is similar to the motivation behind the maximum metric approach, except here we are considering the posterior instead of the metric surface. In this case, it is assumed that the mode of the training set posterior and the mode of the test set posterior will be similar.

Preliminary experiments found that the maximum metric and the maximum a posteriori estimates were the same across a range of models, collections, and metrics in all but a few cases. Even when the two estimates differed, they...
never varied significantly. Therefore, for simplicity, we use the maximum metric approach as the representative point estimate technique throughout the remainder of this paper.

### 3.3 Bayesian Techniques

Note that the maximum metric (maximum a posteriori) technique has the underlying assumption that the metric surface (posterior distribution) is concave (unimodal). The point estimates may yield ineffective results in those cases that do not comply with this assumption. To see why this is the case, consider a query stream with two distinct types of queries. Suppose that, for the first query type, one parameter setting yields the optimal performance, whereas for the second query type, a radically different setting does. Since a point estimate can only choose one of the optimal parameter settings, the resulting model is expected to perform well on one type of query, but poorly on the other type of query, which is undesirable.

Therefore, we investigate two Bayesian techniques that can be used to overcome such phenomenon. In the following approaches, rather than choosing a single estimate to use across all queries, a new parameter is selected for each query (or set of queries). This is done by repeatedly sampling parameters from some underlying distribution.

#### 3.3.1 Uniform Sampling

The most naive approach is to assume a uniform distribution over the entire parameter space. That is, for each query, randomly sample \( \theta \) uniformly, and run the query (or set of queries) using the underlying model. In situations where no training data is available and no prior knowledge on valid parameter settings is available, this may be the only choice available. This approach is expected to perform poorly, especially over models with non-uniform posterior distributions.

#### 3.3.2 Posterior Sampling

Rather than sampling parameters from a uniform distribution, we can sample them directly from the posterior. This is a much more intelligent sampling technique and much more likely to yield good effectiveness. The motivation for such sampling is to overcome the problems involved when the posterior may be multimodal, as discussed previously. In such distributions, sampling can be ‘safer’ and ensure that parameters around each mode are used. If the posterior is highly peaked around a single mode there is a small risk involved with using this approach, as we may sample many suboptimal parameters. Depending on how sensitive the effectiveness metric is around the mode, simply using a point estimate may achieve better effectiveness in this case. These methods are evaluated experimentally in Section 6.

### 3.4 Estimating the Posterior

We have yet to explain where the posterior distribution over the optimal parameters comes from, even though several of the parameter selection approaches described depend on having access to one. There are several options available.

First, a parametric generative model could be used. That is, one can parameterize and estimate both \( P(T|\theta) \) and \( P(\theta) \) to compute the posterior via Bayes’ rule as:

\[
P(\theta|T) = \frac{P(T|\theta)P(\theta)}{\int_\theta P(T|\theta)P(\theta)}
\]

Unfortunately, it may be difficult to properly parameterize \( P(T|\theta) \). In addition, it may be possible to estimate \( P(\theta) \) on a model-by-model or collection-by-collection basis, but is difficult to do in general.

Another option is to directly estimate the posterior using a non-parametric method. In statistics, the bootstrap is a powerful non-parametric technique that can be used to estimate estimator bias, standard error, confidence intervals, and approximate posterior distributions\[2\]. The idea behind the method is, for some plug-in estimator, to start with a sample estimate, and use the original sample to ‘bootstrap’ particular inferences about the population estimate. The actual bootstrapping is accomplished by repeatedly resampling, with replacement from the original sample, and recomputing the estimate on each replicate. Plug-in estimators are a wide class of estimators such that the sample and population estimates are computed in the same manner. The mean, median, and max are all examples of plug-in estimators. Once a large number of estimates have been bootstrapped, we can use the empirical distribution of the bootstrap estimates as an estimate of our posterior.

This bootstrap estimate of the posterior is non-parametric in the sense that there are no underlying distributional assumptions. The larger the original sample size and the number of bootstrap replicates used, the better our estimated
The entropy of a distribution can be thought of as the uncertainty inherent in it. Note that the entropy alone is not a valid indicator of sensitivity. A posterior distribution with a large entropy is not necessarily sensitive, because the metric surface may still be flat. Similarly, if the posterior has low entropy, but the metric surface varies widely over the high confidence parameter values then there exists parameter sensitivity. Therefore, rather than looking at entropy alone, we must also include some notion of the flatness of the metric surface.

### 4.2.2 Spread

In order to measure how flat a distribution over a set of parameter values, we compute the spread of the effectiveness metric. This is computed as:

\[
S = \max_{\theta \in \{\theta : P(\theta | T) > 0\}} m(\theta, T) - \min_{\theta \in \{\theta : P(\theta | T) > 0\}} m(\theta, T)
\]

where \(\{\theta : P(\theta | T) > 0\}\) is the support of \(\theta\).

The spread is only computed over the support of the posterior, rather than over the entire parameter space because it represents the interesting part of the space. This automatically ignores parts of the parameter space that are unlikely to ever be used, and thus should be ignored when computing sensitivity.

The spread, when combined with the entropy, provides a novel, robust way of looking at parameter sensitivity. For example, a model with high entropy, but low spread is more stable than a model with low entropy, but large spread. An ideal model will have low entropy and low spread, which indicates very high confidence over a small, flat range of the parameter space.

### 4.3 Uninformed Sensitivity

The analysis of sensitivity via entropy and spread is one way of looking at sensitivity. It looks at the sensitivity over the support of the posterior, and therefore makes the implicit assumption that a user setting the parameters has access to the same information. This is often not the case, however, as it is often difficult to determine the likely parameter range without some form of relevance judgments.

Therefore, another way of looking at sensitivity is to ask how effective a given model is, given no training data or other information that guides in the selection of an effective parameter setting. In order to measure this, we compute the expected effectiveness of randomly sampling parameters over the entire parameter space, which was discussed in Section 3.3.1. Those models that achieve higher effectiveness in this setting are less sensitive under the “zero information” scenario than those with lower effectiveness.

Sensitivity results are given in Section 6.4.

### 5. PARAMETER GENERALIZATION

Finally, we investigate several aspects of retrieval model generalization. As mentioned in the introduction, generalization is one of the most important aspects of retrieval models. This section ties together ideas from parameter estimation and sensitivity to develop a better understanding of generalization.

### 5.1 Problem Statement

The ultimate goal of parameter selection strategies is to produce a model that generalizes well. A model is said to
generalize well if, when trained on one set of data, remains effective on an unseen test set. A model that is capable of achieving excellent effectiveness on a training set but performs poorly on a test set is of minimal value. Therefore, if some parameter selection method results in effectiveness \( m \), and the optimal effectiveness is \( m^* \), we then compute the following:

\[
G = \frac{m}{m^*}
\]

which we define as the effectiveness ratio. An ideal model, that generalizes perfectly, would achieve an effectiveness ratio of 1. In information retrieval, even a 2-5% change in some measures, such as mean average precision, can be a statistically significant difference, and therefore effectiveness ratios below 0.90 indicate the model’s inability to generalize can severely hinder its effectiveness. Most reasonable retrieval models will have an effectiveness ratio greater than 0.95.

5.2 Generalization Types

We are particularly interested in intracollection and intercollection generalization, which are two different ways of measuring the generalization properties of a model, which we now describe.

5.2.1 Intracollection Generalization

Intracollection generalization deals with how well a model trained on a set of topics from some collection generalizes to another set of topics on that same collection. This is a common setting in TREC evaluations, where collections are often reused from year to year, and systems are typically trained on the topics from the previous year(s).

5.2.2 Intercollection Generalization

The other type of generalization we consider is intercollection generalization. This type of generalization measures how well a model trained on a topic set from one collection generalizes to a different topic set on a different collection. This is a practical scenario for ‘off the shelf’ retrieval systems that may be used across a wide range of different collections. It is unlikely that the end users of these systems will be willing or able to provide training data to the system, and therefore the system must be shipped with a very solid set of pre-tuned, highly generalizable parameters.

6. EXPERIMENTAL EVALUATION

We now describe the experiments carried out to analyze the tools and methods developed up until this point. Although focused entirely on title-only ad hoc information retrieval experiments, our results and analysis are general enough to provide insights into a variety of retrieval tasks.

6.1 Data

Experiments are carried out over a set of five standard TREC collections, varying in size and type. Three collections are moderately sized newswire collections, while the other two are large web collections. For each collection, a set of TREC ad hoc, title-only topics are used, with a set devoted to training, and another for testing purposes. Summary statistics of the collections are given in Table 1.

6.2 Models

In this section we give a brief overview of the retrieval models used in our experiments. They consist of two classical models (language modeling, BM25) and recent variants of each (dependence model, F2EXP). All models are implemented using a modified version of the Indri [11] search engine. All documents are processed using the Porter stemmer and a standard stopword list.

6.2.1 Language Modeling (Dirichlet Model)

Language modeling is a robust statistical framework for modeling documents and queries [1]. There are various ways to use the framework to rank documents, such as query likelihood, document likelihood, and KL-divergence. In addition, there many ways to model and smooth documents [12]. In this work, we choose to model documents as unigram multinomial language models, smoothed with Dirichlet smoothing, and rank documents according to query likelihood. These assumptions result in the following ranking formula:

\[
P(Q|D) = \prod_{w \in Q} \left( \frac{tf_w, D + \mu}{|D| + \mu} \right)^{tf_w, Q}
\]

where \( tf_w, Q \) and \( tf_w, D \) are the number of times term \( w \) occurs in the query and document, respectively, \( c_w \) is the number of times \( w \) occurs in the entire collection, \( |D| \) is the length of the document, and \( |C| \) is the length of the collection, and \( \mu \) is the single free parameter, which is used to control the amount of smoothing.

6.2.2 Dependence Model

Metzler and Croft recently proposed a model that is loosely a variant of language modeling which incorporates term dependencies [7]. The model proved to be consistently more effective than language modeling, especially on larger web collections. The model considers single term, ordered window, and unordered window features. Given a query, these features are automatically extracted in a spirit similar to that done by Mishne and de Rijke [8]. Full details are omitted due to space constraints. However, once the features are extracted, an Indri structured query is formed, which does a weighted combination of each feature type [6]. For example, given the query scottish highland games, the following query would be generated:

\[
\#weight( \lambda_T \ #combine( \text{scottish highland games} ) \\
\lambda_O \ #combine( \#1 \text{highland games} ) \\
\#1( \text{scottish highland} ) \\
\#1( \text{scottish highland games} ) ) \\
\lambda_U \ #combine( \#uw8( \text{highland games} ) \\
\#uw8( \text{scottish games} ) \\
\#uw8( \text{scottish highland} ) \\
\#uw12( \text{scottish highland games} ) )
\]

where \( \lambda_T, \lambda_O, \text{and } \lambda_U \) are free parameters corresponding to single term features, ordered term features, and unordered term features respectively. Documents are ranked by running the query using the Indri search engine.

Table 1: Overview of collections and topics used.
6.2.3 BM25

The BM25 retrieval model, inspired by the 2-Poisson model, is effective and widely used [10]. It is composed of a standard idf and specialized tf component. Documents are ranked according to:

\[ S = \sum_{w \in Q \cap D} tf_{w,Q} \frac{(k_1 + 1)tf_{w,D}}{k_1 (1 - b + b \frac{df_w}{df_{\text{avg}}})} \log \frac{N - df_w + 0.5}{df_w + 0.5} \]

where \( |D|_{\text{avg}} \) is the average document length, \( df_w \) is the number of documents term \( w \) occurs in, \( N \) is the number of documents in the collection, and \( k_1 \) and \( b \) are the model parameters.

6.2.4 F2EXP

As mentioned in the Section 2, Fang and Zhai [4] proposed a variant of BM25 called F2EXP that was shown to be at least as effective as BM25 while being less sensitive to parameter settings over a range of collections. Using the model, documents are ranked according to:

\[ S = \sum_{w \in Q \cap D} tf_{w,Q} \frac{tf_{w,D}}{tf_{w,D} + s + s \frac{|D|}{|D|_{\text{avg}}}} \left( \frac{N + 1}{df_w} \right)^k \]

where \( s \) and \( k \) are the two free parameters.

6.3 Estimation Results

The goal of the estimation experiments is to determine if one parameter selection strategy dominates the others. It was argued that the maximum metric method is likely to be the best choice for unimodal distributions and that the posterior sampling method could lend itself to multimodal or unstable distributions. Results were performed for each of the retrieval models, on each of the collections, using both the maximum metric and posterior sampling method. The maximum metric estimate and posterior distribution were estimated on the training topics and evaluated on the test topics. The results are given in Table 2.

We note that the posterior sampling results are asymptotic results. That is, they work under the assumption that the test set was averaged over an infinitely large number of runs. Since parameters are randomly chosen, the actual effectiveness of individual runs will vary slightly.

As the results show, neither method clearly dominates. Both methods, in fact, work well for all models/collection pairs, with negligible differences between the methods in most cases.

There are several interesting trends in the data, however. For the w110g collection, it is almost always better to use the sampling method. Figure 1 gives an indication of why this may be the case. The posterior distribution estimated using the Dirichlet model has several modes, which indicates that there may actually be multiple optimal parameter values that are appropriate, rather than a single, fixed value.

The benefits of the fixed max metric approach include the ease of implementation, as no posterior needs to be estimated, and a deterministic set of retrieval results. On the other hand, the posterior sampling method has the advantage that it can handle unstable and multimodal parameter distributions better and also, in general, has lower variance in effectiveness across query sets. Although the differences in effectiveness achieved here are small, there may exist scenarios where they vastly differ, depending on characteristics of the posterior distribution. Therefore, the best parameter selection strategy is largely dependent on the underlying behavior of the model, retrieval task, and query stream.

6.4 Sensitivity Results

6.4.1 Informed Sensitivity

As discussed in Section 4, both the entropy and spread of the posterior distribution must be considered together to determine the sensitivity of a model. Thus, for the sensitivity experiments, the entropy and spread for each model over every collection was first computed. Then, for four of the collections, the models were plotted with regard to these measures. These plots are given in Figure 2.

In terms of spread, the results indicate that the F2EXP and Dirichlet models consistently produce flatter ranges of mean average precision values over BM25 and the dependence model. This indicates these models are less sensitive in this regard. This result agrees with previous work that found the F2EXP model to be less sensitive, in terms of flatness, than the BM25 model [3, 4]. We also note that the spread on the web collections is larger than the spread on the newswire collections. This is likely caused by the large, noisy nature of such collections, where there is more room for large changes in effectiveness over smaller, less noisy collection, such as the newswire ones.

When looking at entropy alone, there are fewer conclusions that can be drawn. However, when coupled with spread, we see that the F2EXP and Dirichlet models tend to have the smallest combination of spread and entropy, which indicates that they models have a small, stable range of valid parameters compared to BM25 and the dependence model. This further supports our parameter selection hypothesis that Bayesian sampling techniques may be more fruitful for less stable distributions. As Table 2 shows, the sampling selection technique is the best choice more for both BM25 and the dependence model than the other two models.

Therefore, our results indicate that, in terms of sensitivity, the F2EXP and Dirichlet models are the least sensitive models, and that slight variations in their parameter settings are less likely to produce drastic changes in effectiveness. In addition, the results indicate that both of these models have rather focused (low entropy) posterior distributions. However, whether or not sensitivity corresponds to generalization properties of a model has yet to be explored.

We return to this issue shortly.

6.4.2 Uninformed Sensitivity

To evaluate the models in terms of uninformed sensitivity, effectiveness ratios were computed using the uniform sampling parameter estimation technique discussed in Section 3.3.1. Values were uniformly sampled for all free parameters over a reasonably wide range of values for each model based on numbers quoted in previous studies. For example, for the F2EXP model, the parameter values \( s \) and \( k \) were uniformly sampled from the range 0 to 1 [4]. The results are given in Table 3.

As we see from the table, the Dirichlet model is by far the most stable, always achieving an effectiveness ratio of at
Table 2: Comparison between maximum metric (point) and posterior sampling (sample) parameter selection techniques. Bold values indicate best selection strategy for each model/collection pair. Effectiveness is measured by mean average precision.

<table>
<thead>
<tr>
<th>Model</th>
<th>Dirichlet</th>
<th>BM25</th>
<th>F2EXP</th>
<th>DM</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Point</td>
<td>Sample</td>
<td>Point</td>
<td>Sample</td>
</tr>
<tr>
<td>ap</td>
<td>0.2074</td>
<td>0.2076</td>
<td>0.2143</td>
<td>0.2142</td>
</tr>
<tr>
<td>wsj</td>
<td>0.3258</td>
<td>0.3260</td>
<td>0.3331</td>
<td>0.3315</td>
</tr>
<tr>
<td>robust04</td>
<td>0.2020</td>
<td>0.2012</td>
<td>0.2805</td>
<td>0.2805</td>
</tr>
<tr>
<td>wt10g</td>
<td>0.1862</td>
<td>0.1909</td>
<td>0.1960</td>
<td>0.1960</td>
</tr>
<tr>
<td>gov2</td>
<td>0.3234</td>
<td>0.3230</td>
<td>0.3335</td>
<td>0.3311</td>
</tr>
</tbody>
</table>

Figure 2: Sensitivity plots over two newswire collections (ap, robust04) and two web collections (wt10g, gov2). Models are plotted based on their entropy and spread on the training set of topics.

Table 3: Uninformed sensitivity results. Values are the expected effectiveness ratios achieved on the test topics by uniformly choosing parameters over entire parameter space.

<table>
<thead>
<tr>
<th>Model</th>
<th>Dirichlet</th>
<th>BM25</th>
<th>F2EXP</th>
<th>DM</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>ap</td>
<td>wsj</td>
<td>robust04</td>
<td>wt10g</td>
</tr>
<tr>
<td></td>
<td>99.19</td>
<td>95.62</td>
<td>95.62</td>
<td>97.57</td>
</tr>
<tr>
<td></td>
<td>94.39</td>
<td>92.01</td>
<td>91.30</td>
<td>85.45</td>
</tr>
<tr>
<td></td>
<td>90.62</td>
<td>91.12</td>
<td>90.35</td>
<td>92.05</td>
</tr>
<tr>
<td></td>
<td>91.19</td>
<td>91.04</td>
<td>92.16</td>
<td>89.29</td>
</tr>
</tbody>
</table>

Table 5: Intercollection generalization results. Table includes mean average precision effectiveness ratios across all possible train/test splits using the F2EXP model.

Table 6: Intercollection generalization results. Table includes mean average precision effectiveness ratios across all possible train/test splits using the dependence model.

6.5 Generalization Results

6.5.1 Intracollection

For the intracollection experiments, parameter settings are chosen using the max metric method on the training set for each collection. Models are evaluated using the effectiveness ratio on the test set. The metric used is mean average precision. The results are given in Table 4.

The table lists both the optimal mean average precision value achievable and the effectiveness ratio of the trained parameter setting. A model with perfect intracollection generalization would have an effectiveness ratio of 100. The results indicate that all of the models do a relatively good job of generalizing, with average effectiveness ratios well above 98%. We note that the F2EXP model tends to generalize better within newswire collections, while the dependence model generalizes better for web collections. The BM25 model, however, has the best average effectiveness ratio, which indicates its parameters do a particularly good job of capturing collection-dependent characteristics, rather than topic set-specific ones.

6.5.2 Intercollection

When measuring intercollection generalization, we chose a parameter setting for a collection using the max metric method on the combined set of training and test topics. We then computed the effectiveness ratio of using this parameter setting across the other collections. The results are given for the F2EXP and dependence models in Tables 5 and 6, respectively.
As we see from the table, the cross-collection effectiveness ratios for the dependence model are higher across for every training/test set pair, with very few exceptions. In fact, on average, the dependence model comes within 1% of the optimal setting regardless of which collection is being trained on, whereas the F2EXP model only comes within 4% of the optimal on average. The Dirichlet and BM25 models (not shown) have average effectiveness ratios of 98.9% and 96.9%, respectively. Therefore, the dependence model and Dirichlet models are more robust when it comes to cross-collection generalization and make them good candidates for “out of the box” implementations that require a single parameter setting to work well across a wide range of collections. It is worth noting that the dependence model, on average, outperforms the Dirichlet model by 6.4%, BM25 by 6.1%, and F2EXP by 8.8% in terms of mean average precision across collections, which further adds to its applicability.

Finally, we note that there is a certain disconnect between sensitivity and generalization. Models that are less sensitive are not necessarily those that generalize the best. This is mainly caused by the characteristics of the posterior distribution. If the distribution changes across collections, the dependence model comes within 1% of the optimal, the Dirichlet model performs the Dirichlet model by 6.4%, BM25 by 6.1%, and F2EXP by 8.8% in terms of mean average precision across collections, which further adds to its applicability.

7. CONCLUSIONS

In this work we developed and evaluated a set of statistically motivated tools for choosing parameters, measuring sensitivity, and studying generalization properties of parameterized retrieval models.

First, we showed that the maximum metric estimate is not always the optimal parameter selection strategy and that a Bayesian sampling technique may be more appropriate in the case of multimodal or unstable posterior distributions. Next, two novel measures were introduced to measure different aspects of sensitivity. The results showed that the F2EXP and Dirichlet models were the most stable according to these measures, having relatively flat, focused metric surfaces.

Finally, we looked at intracollection and intercollection generalization and showed that there was a disconnect between sensitivity and generalization. That is, simply because a model is stable does not mean it will generalize well. We showed that the KL-divergence between posteriors is strongly indicative of how well an estimate will generalize. Despite the lack of stability with regard to entropy and spread, the dependence model was found to generalize better than the stable F2EXP model, especially across collections.

Acknowledgments

This work was supported in part by the CIIR, in part by NSF grant #CNS-0454018, in part by ARDA and NSF grant #CCF-0205575, and in part by NSF grant #HS-0527159. Any opinions, findings and conclusions or recommendations expressed in this material are the author’s and do not necessarily reflect those of the sponsor.

8. REFERENCES